A Reformulation of Libertarian Paternalism

Guilhem Lecouteux

ECOLE POLYTECHNIQUE

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

A REFORMULATION OF LIBERTARIAN PATERNALISM

Guilhem LECOUTEUX

July 2013

Cahier n° 2013-18

DEPARTEMENT D'ECONOMIE
Route de Saclay
91128 PALAISEAU CEDEX
(33) 1 69333033
http://www.economie.polytechnique.edu/
mailto:chantal.poujouly@polytechnique.edu

A REFORMULATION OF LIBERTARIAN PATERNALISM

Guilhem Lecouteux[*]

July 2013

Abstract

Conventional normative economics is built on the assumption that people act as if seeking to satisfy coherent and *a priori* preferences. This model has however been challenged by many empirical works highlighting the existence of systematic deviations from the behaviour predicted by the neoclassical theory. The development of behavioural economics therefore questions the validity of the results developed by normative economists. Reconciling behavioural and normative economics needs in particular a clarification of the normative content of economic prescriptions, since it appears that the assumption of rational preferences enabled economists to overstep this question, the different interpretations of the current normative criterion of preference-satisfaction leading *in fine* to the same prescriptions. In this paper, we want to highlight that libertarian paternalism is probably the most natural solution to the reconciliation problem for neoclassical economists, since its current formulation relies on the existence of a rational *homo œconomicus* trapped within each individual. We can however find within the current formulation of libertarian paternalism the same difficulties than the ones of Pareto's theory of the *homo œconomicus*. We therefore suggest a reformulation of libertarian paternalism based on a normative criterion of individual autonomy rather than preference-satisfaction, and defend its relevance in the specific context of common-pool resources, by showing that the normative prescriptions generated by our principle of individual autonomy present strong similarities with the institutional design principles of Ostrom (1990) enabling a sustainable management of common-pool resources.

Keywords: welfare economics, libertarian paternalism, *homo œconomicus*, nudge, autonomy, common-pool resources.

JEL classification: B13, D03, D60, Q20.

[*] École Polytechnique, Laboratoire d'économétrie PREG-CECO (CNRS UMR 7176), 91128 Palaiseau, France.
Email: guilhem.lecouteux@polytechnique.edu

1. Introduction

The recent development of behavioural economics produced a large literature showing the existence of numerous and systematic inconsistencies in individual preferences (see for instance Kahneman and Tversky 2000 and Camerer 2003). Those evidence raise serious issues for welfare economics, since its main results are build on the assumptions that the individuals are rational and act as if seeking to satisfy coherent preferences. Furthermore, this also questions standard methods of welfare analysis such as cost-benefit analysis, which assumes the stability of individual preferences and the possibility of assessing individual welfare through the satisfaction of individual preferences. The difficulty of reconciling behavioural and normative economics is that the normative criterion used by welfare economists is the satisfaction of individual preferences: McQuillin and Sugden (2012) argue that this criterion can be interpreted in three conceptually different ways – as an evaluation in terms of happiness, self-assessed well-being or freedom – which lead to the same prescriptions as long as the preferences of the individuals are coherent. However, since it empirically appears that the preferences of the individuals are generally incoherent, the different interpretations of the preference-satisfaction criterion do not lead to the same prescriptions any more: economists must therefore choose one of these interpretations in order to clarify the normative content of their prescriptions. In the first case, economists evaluate the soundness of a policy according to the welfare it generates, *objectively* defined in a mental-state perspective. According to this approach, some mental states are intrinsically good, such as pleasure or happiness – in the spirit of Bentham's hedonism (Kahneman *et al.* 1997). In the second

case, the evaluation is in terms of a *subjective* notion of welfare, defined as what the individual values as preferable to herself if she had "complete information, unlimited cognitive abilities, and no lack of self-control" (Sunstein and Thaler 2003, 1162). This approach was introduced by Sunstein and Thaler (2003) and Camerer *et al.* (2003), and popularized by Thaler and Sunstein (2008) thanks to their book *Nudge: Improving Decisions about Health, Wealth, and Happiness*. The proponents of this criterion defend a "soft paternalism" – "libertarian paternalism" for Sunstein and Thaler and "asymmetric paternalism" for Camerer *et al.* – and justify paternalistic interventions if they correct the possible decision flaws of the individuals without coercing their choices. In the third case, the evaluation does not focus any more on the effective choices of the individuals, but on the opportunity they had making choices among a wide range of alternatives. This argument has been developed by Sugden (2004, 2007), who suggests using a normative criterion of opportunity rather than preference-satisfaction.

The second interpretation in terms of self-assessed well-being derives from the standard model of the individual used in normative economics: the individual is indeed conceived as a rational entity "trapped" in a non rational one, the latter offering a biased perception of the world to the former, who is then unable to properly satisfy her preferences. The aim of the social planner is then to help the individual to correct her eventual decision flaws, and therefore to become rational and to satisfy her "true" preferences. There exists here a duality between the actual preferences of the individual – the relation that determines her effective choice – and her true preferences, i.e. the relation that would have determined her choice if she was perfectly rational. Sunstein and Thaler therefore defend the idea that, since the individuals can make mistakes (in the sense that they do

not satisfy their true preferences), they should be nudged by the individual in charge of the design of the situation of choice – the "choice architect" – so that they satisfy *in fine* their true preferences.

Libertarian paternalism has been presented as "the real Third Way" (Thaler and Sunstein 2008, 252-253), and many behavioural economists showed a great interest in soft paternalism as a solution to the reconciliation problem (see for instance Kahneman (2011, 412-415) who describes *Nudge* as a "bible for behavioral economics"). In order to constitute a real alternative to paternalism and libertarianism, we suggest that two conditions need to be fulfilled:

(i) Subjectivity condition: it is possible to impartially define the true and subjective preferences of the individuals, and to implement a choice architecture such that they satisfy those true preferences.

(ii) Free-choice condition: when nudged by a choice architect, the individuals are still able to make free choices.

Under those conditions, libertarian paternalism can indeed be considered as a real alternative to paternalism, since the planner respects the subjectivity of the individuals (and does not try to implement what she thinks is the "right" choice from her own point of view) and does not coerce them. In this paper, we suggest discussing those two conditions. We firstly show that the subjectivity condition can be verified if and only if we accept the empirical validity of the neoclassical model of individual behaviour. This conclusion is obviously in contradiction with the central objective of behavioural

economists which is to build theories of individual behaviour grounded on psychology rather than on *a priori* principles of rational choice. We then show that, if we accept Suntein and Thaler's conception of the individual, the free-choice condition can be verified if and only if the individuals are not sensible to nudges, i.e. if and only if libertarian paternalism cannot be efficient. We indeed argue that the freedom of choice should be evaluated through the set of options within which an individual is actually able to make her choice, and not the set within which a rational individual would be able to make her choice. We then suggest a possible reformulation of libertarian paternalism consistent with its initial spirit and respectful of the subjectivity and the free-choice conditions.

This paper is organized as follows. In section 2, we show that the reasoning of Sunstein and Thaler presents strong similarities with the methodological approach of Pareto and his theory of the *homo œconomicus*. Thanks to this parallel, we show in section 3 the difficulties of isolating the true preferences of the individual, and therefore of respecting the subjectivity condition. In section 4, we suggest that the mere efficiency of libertarian paternalism implies that the free-choice condition cannot be verified within Sunstein and Thaler's framework. We then present our reformulation of libertarian paternalism, grounded on a normative criterion of individual autonomy rather than preference-satisfaction. In section 5, we justify our normative criterion in the context of the management of common-pool resources, by showing that the institutional design principles of Ostrom (1990) that characterize robust institutions for managing common-pool resources perfectly correspond to the policy that a government guided by a criterion of individual autonomy would try to implement.

2. Libertarian paternalism and logical actions

The fundamental argument that justifies libertarian paternalism is the idea that paternalism is unavoidable. People are subject to numerous decision biases, such as optimism and overconfidence (Sunstein 1998), loss aversion (Kahneman and Tversky 1979) and status quo biases (Samuelson and Zeckhauser 1988). They are therefore "Humans" – *homo sapiens* – and not merely "Econs" – *homo œconomicus* (Thaler and Sunstein, 2008, 6). Since people are sensible to framing effects, the choice architect can slightly influence their choices. Sunstein and Thaler illustrate this phenomenon with the management of a cafeteria. They imagine the work of Carolyn, in charge of the location of the different food items (2003, 175). Carolyn knows how the organisation of her cafeteria can change the choices of the individuals. Since the choice architecture is not neutral, the choice architect can influence the choices of the decision makers in order to satisfy a specific objective: paternalism is therefore unavoidable. It becomes then imperative to determine what objective the choice architect should pursue. Sunstein and Thaler argue that the most legitimate objective is to help people to improve their well-being, as judged by themselves (2008, 5). The logic of libertarian paternalism is therefore that people aim to choose the options that will make them better off, but that – due to human fallibility – they often make non rational choices, in the sense that they miss their objective. Since a choice architect has the possibility of slightly influencing people's choices, she should nudge them so that they achieve *in fine* this objective. Since this approach is not coercive (the individuals are indeed not forced to choose the option the choice architect wants them to choose), Sunstein and Thaler argue that the freedom of choice of the decision makers is preserved. Furthermore, it seems very appealing to

nudge people towards what they would have chosen if they "had complete information, unlimited cognitive abilities, and no lack of self-control", since people would probably agree with this objective if they were aware that they are nudged by the choice architect.

However, despite their references to behavioural considerations and their reject of the model of the *homo œconomicus* as a descriptive model of human behaviour, Sunstein and Thaler's reasoning is paradoxically grounded on this same model. Their reasoning is indeed the following: people have subjective preferences but, since they are Humans and not Econs, they can miss to satisfy them; as a planner (a choice architect), we should therefore help them to satisfy their subjective preferences; to do so, we must then identify what their "true" preferences are, such that we will be able to implement the adequate choice architecture. Those true preferences – by opposition to the effective preferences of the individual that determine her effective choice – correspond to the relation that would have determined her choice if she was perfectly rational, i.e. if she was an Econ and not a Human: isolating a rational component (the true preferences) within the effective preferences precisely corresponds to the methodological approach suggested by Pareto and his notion of the *homo œconomicus*.

Pareto (1909, 1916) grounded his sociology on a classification of human actions according to two criteria, whether the action is undertaken (subjectively) in order to satisfy a given purpose or not, and whether this action is objectively appropriate towards the subjective purpose of the individual. Pareto then defined logical actions as the set of actions for which the individual has an intention of satisfying a purpose, when this action is objectively appropriate towards this purpose. When the objective purpose (i.e. what the

individual will really achieve) differs from the subjective purpose (what the individual tries to achieve), the actions are non-logical, and can be classified into different categories, according to the nature of the difference between the objective and the subjective purpose ([1916] 1936, §151). Therefore logical actions "logically conjoin means to ends not only from the standpoint of the subject performing them, but from the standpoint of other persons who have a more extensive knowledge" ([1916] 1936, §150). Pareto then suggested a methodological reductionism and claim that the individual can be studied as the aggregation of different kind of *homines*, according to the nature of the action ([1909] 1971, Chap.1). In particular, he defined a *homo œconomicus* as the "dimension" of human being which deals with logical actions: the *homo œconomicus* therefore models the behaviour of an individual who knows what her objectives are and how to satisfy them.

We can underline several similarities between Pareto's approach and the model of human behaviour that supports libertarian paternalism. Sunstein and Thaler assume that individuals want to improve their well-being, as judged by themselves: we find here the first criterion of a logical action, the intention of satisfying a subjective purpose. They then suggest that individuals often make bad choices (either due to information issues or to a lack of self-control), and that the choice architect can help them to achieve their purpose. To reuse the definition of Pareto, we can say that individuals often make actions which logically conjoin means to ends from their own standpoint, but not from the standpoint of the choice architect, a person who is supposed to have a more extensive knowledge. Libertarian paternalism is therefore grounded on the will of creating a choice architecture such that the individuals perform *in fine* logical actions. However, unlike

Pareto who only considered the *homo œconomicus* as a descriptive model of human behaviour (empirically valid in a few settings such as repeated markets (Plott 1996)) and a useful abstraction for the study of markets (Pareto [1909] 1971, Ch.3, §65-66 and §87)), Sunstein and Thaler describes this rational entity as a normative model of human behaviour, i.e. what the individuals would like to be while making choices. They implicitly assume that being rational improves individual well-being, and therefore that it constitutes a valid normative model of behaviour. Indeed, assuming that an individual would accept to be nudged if it enables her to improve her well-being implies that, in a state of perfect rationality, her preferences are "better" than her effective preferences. There however exist several games for which this condition is not true, i.e. games for which rationality is self-defeating. In those games, satisfying preferences different from one's true preferences can improve the well-being of the decision-maker[1]: this implies that people would not necessarily agree to become rational, and therefore that – on a purely technical level – the *homo œconomicus* is not necessarily an acceptable normative model of behaviour.

Furthermore, it seems a bit ironic that the proponents of libertarian paternalism – who are also behavioural economists – criticized the model of a rational *homo œconomicus* as a description of actual human behaviour, but want to define this same *homo œconomicus* as the ideal the individuals want to be. Indeed, a normative model of human behaviour should present some empirical relevance, since the choice architect must predict the behaviour of the individual if she was a *homo œconomicus*, which presupposes that this situation is possible (or at least credible). This model of behaviour is however quite

[1] It is for instance the case of coordination games (such as the Hi-Lo game, in which rational players are unable to coordinate themselves), or of commitment games, such as the Toxin Puzzle (Kavka 1983).

implausible: it necessitates the existence of "true" preferences, *a priori* and immutable. The decision making process is then conceptualized as follows: an individual has a clear set of ends (subjective and given *a priori*), a set of means at her disposal, and she chooses the means that best satisfy her ends. The individual is then nothing more than a computer, "fitting given means to given ends" (Georgescu-Roegen 1971, 343). It can here seem a bit surprising that, as behavioural economists, the proponents of libertarian paternalism accept such a model based on principles of rational choice and free from any psychology. However, despite those criticisms, it is still possible to assume that being rational constitutes an acceptable normative model of behaviour, and that each individual actually presents "true" preferences. Although the representation of the individual as a rational *homo œconomicus* trapped in a non rational body who tries to satisfy some true preferences is questionable, it is still possible to accept libertarian paternalism as a Real Third Way if it is possible to isolate those true preferences and to implement the *ad hoc* choice architecture, i.e. if the subjectivity condition can be verified.

3. Isolating the true preferences

Suppose now that the *homo œconomicus* constitutes an acceptable normative model of human behaviour, and that each individual has true preferences. The planner needs now to identify those true preferences, so that she will be able to design the adequate choice architecture. A first solution would be to deduce them from the effective preferences of the individual. This approach is for instance endorsed by Bleichrodt *et al.* (2001), who assume that the individuals present loss aversion, and try then to deduce the true and unbiased preferences of the individuals from their revealed preferences. However, we

cannot know *a priori* the list of the different biases that influence individuals' decisions. In the case of Bleichrodt *et al.*, it can be doubtful to assume that the individuals only present loss aversion, since it is not necessarily their only bias, and it is not certain that they really suffer from loss aversion. Knowing the actual preferences is therefore probably insufficient to obtain the true preferences, since there is an issue of identification within the determinants of behaviour between the true preferences and the possible decision flaws that affect the choices of the individual. It becomes therefore necessary to implement an impartial mechanism able to reveal the true and unbiased preferences of the individuals.

Several authors – including Pareto (1909, Chap.3, §1) – suggest that the discovery of the true preferences is the product of learning thanks to the repetition of the situation of choice. Binmore stresses for instance that people tends to behave according to the rational choice theory (and therefore to act like the *homo œconomicus*) if the problem "seems simple to the subjects", the "incentives provided are 'adequate'", and the "time allowed for trial-and-error adjustment is 'sufficient'" (Binmore 1999, F17). A choice architect can therefore perform repeated experiments involving the individuals in order to deduce their true preferences. However, since the true preferences of the individual are defined subjectively, it is not possible to know at which time the individual is satisfying them: there is indeed still an issue of identification, since it is maybe not possible to distinguish between the true preferences and a systematic decision flaw. There is therefore here the temptation of defining *objectively* the ends of the individual (for instance as selfish ones), and to consider that the individual has achieved this state of rationality and performs logical actions if and only if she is satisfying the preferences expected by the

experimenter. This is for instance what Binmore is doing in his analysis of the ultimatum game:

> "Novices offer a fair amount because this is what their currently operative social norm recommends. Novices who are offered unfairly small amounts are programmed to feel resentful and so want to punish the proposer by refusing. But this behaviour changes over time as people dimly perceive that the norm they are using is not adapted to the problem with which they are faced. In the Ultimatum Game, people learn that it does not make much sense to get angry if offered too little, but the mavericks who initially make small offers learn much faster that it does not make sense to demand too much if one is nearly always refused." (Binmore 1999, F22)

Binmore considers that the individuals tend to a payoff maximizing behaviour (since "people learn that it does not make much sense to get angry if offered too little"), but we cannot directly observe it since "the mavericks who initially make small offers learn *much faster* that it does not make sense to demand too much if one is nearly always refused". There is therefore here the implicit assumption that the individuals respect a social norm because they want to maximize their payoff, although we cannot know what the motives of the individuals are: we can for instance consider that an individual respects a specific norm by conformism (see for instance the famous experiment of Asch 1955), or – as suggested by Binmore himself (F19) – that the subjects want to achieve what they

perceive as the experimenter's objective[2], since this one can be perceived as an authoritative figure (Milgram 1975).

Since we cannot make a clear distinction between the ends of the individual and the different factors that can influence her decision, it seems quite difficult to design an experiment for which "the time allowed for trial-and-error adjustment is 'sufficient'". An apparent stable behaviour can indeed correspond to the pursuit of a specific end plus a systematic decision flaw. In the previous example, we can for instance assume that the true objective of an individual is to offer and accept only equal shares, but she prefers to follow an unfair rule that was implemented during the experiment by conformism. Although the actual behaviour of the individual is well predicted by the theory according to which the individuals want to maximize their payoff, the underlying reasons of her choice are not the simple maximization of her payoff.

This raises considerable practical issues for libertarian paternalism. Since the choice architect tries to implement a choice architecture that will lead the individuals to perform logical actions, it is essential to clearly define what would be the choice of the individuals if they were able to perform logical actions. The crucial issue is that logical actions are defined by conditions such that "sufficient" repetitions and "adequate" incentives: the qualification of "logical" for a specific action is therefore subject to the personal interpretation of the experimenter. It seems therefore implausible to implement an impartial procedure which could isolate the true preferences from the actual preferences

[2] In the case of experimental economics, we can typically observe that many participants are students in economics, who can therefore be aware of the phenomenon the experimenter wants to study.

of the individual. The existence of a stable behaviour can indeed correspond to the persistence of a decision flaw, and not necessarily to the discovery of the only true preferences. It means in particular that the choice architect will be forced to guess what the preferences of the individuals are, and is therefore not certain to respect their true preferences.

Suppose nevertheless that the choice architect is able to implement an impartial mechanism that enables her to determine the true preferences of the individuals. The next difficulty of libertarian paternalism is the mere identity of the choice architect: it is indeed assumed that there exists a benevolent and omnipotent planner, who has the will and the ability to implement the choice architecture that will satisfy the true preferences of the individuals. Indeed, as illustrated by Sunstein and Thaler with the example of the cafeteria, the choice architect Carolyn is a Human and not an Econ (more generally, they stress that many real people are choice architects, such as a doctor who must describe different possible treatments available to a patient or a parent describing the possible educational options to her son or daughter (2008, 3)). She also makes choices – since she has in charge the implementation of a choice architecture – and can make mistakes while pursuing her benevolent objective. Furthermore, as a Human, she is not necessarily philanthropic and incorruptible, and can use her position of choice architect to obtain a personal advantage[3]. The logic of libertarian paternalism relies on the existence of an Econ, able to nudge the individuals, but who does not need to be nudged.

[3] Sunstein and Thaler list different possible objectives for the manager of a cafeteria, and consider in particular the objective of "[maximizing] the sales of the items from the suppliers that are willing

Under the conditions that each individual presents *a priori* preferences, it seems delicate for a choice architect to distinguish between a systematic bias and the true preferences of the individual. This difficulty is related to Pareto's notion of logical action, and in particular the difficulty of objectively defining the conditions under which the individuals can perform logical actions. In this case, the planner is forced to decide what the true preferences of the individuals are. Furthermore, once the planner knows her objective, it is assumed that she does not need to be nudged nor monitored by another planner: this conception of the benevolent and rational planner is hardly acceptable, since libertarian paternalism is grounded on the idea that individuals – and therefore choice architects too – need to be nudged. The subjectivity condition cannot probably be verified: identifying the self-assessed well-being of the individuals necessitates indeed the existence of a true and rational self within each individual – the *homo œconomicus* – whose preferences can be discovered by a benevolent planner, who must be an Econ and not a Human. Since there exists an issue of identification between the true preferences and systematic biases, the objective the planner wants to implement is possibly not the objective the individual would have pursue if she was rational, with perfect information and without lack of self-control.

4. Nudges and freedom of choice

to offer the larger bribes" (2008, 2). They simply dismiss this option by assuming that "Carolyn is honorable and honest" (2008, 3).

Although we show in the previous section that isolating the true preferences of the individual – and then implementing the adequate choice architecture – is probably not possible, we can still assume that the choice architect manages *in fine* to discover them (by implementing an impartial procedure which enables her to isolate the true preferences from the actual ones with certainty), or that what she thinks to be the true preferences is relatively close to the true preferences of the individual. Under those conditions, it will be possible to interpret libertarian paternalism as an alternative to paternalism if the free-choice condition is verified, i.e. if an individual who is nudged towards her true preferences is still able to make free choices. The central issue here is to precisely define what the freedom of choice is: this question is indeed tightly intertwined with the issue of the definition of the self. We suggest defining the freedom of choice as the ability to choose without being coerced by elements external to our self, whether or not the choices of the self are determined or predictable. We define an individual able to make free choices as an autonomous individual.

If we consider the model of the individual retained by the proponents of libertarian paternalism, then the self is the *homo œconomicus*, and individual choices can be influenced by external factors such as psychological biases and framing effects. We suggest now showing that this conception of the individual – necessary to the respect of the subjectivity condition and the discovery of the true preferences – implies that the free-choice condition cannot be verified. Indeed, Sunstein and Thaler evaluate the freedom of choice through the set of actions an individual has at her disposal – since they consider that nudges, when they are “easy to avoid”, does not restrict the freedom of choice of the individuals (2008, 6) – and not through the set of actions within which an individual is

actually able to choose her action (i.e. within which she can make free choices). Our point is that it not self-evident to argue that an individual whose choices are conditioned by frames can make free choices. Indeed, Suntein and Thaler emphasize that nudges are not coercive and do not limit the freedom of choice of the individuals, since they can still choose another option that the one wanted by the choice architect *if they want it*. The central difficulty of this argument is that real individuals, unlike the *homo œconomicus*, are generally not able to choose what they "really" want: this is precisely the reason why, according to Sunstein and Thaler, people should be nudged. Suppose for instance that behavioural economists discover a specific frame such that, when the individual is not aware that she is subject to framing effects, she will systematically choose a specific option (a kind of default option for instance). It implies that the choice of the real individual is determined by the only choice architecture: does the existence of alternative options increase the freedom of choice of the individual in this situation? Since her choice will always be the default option, unless someone told her that she is subject to framing effects, the set of options within which she is actually able to select her choice does not extend: this situation is therefore equivalent, in terms of freedom of choice, to a situation in which the only available action is the one chosen by the choice architect. The only difference is that the individual have the illusion of having a greater freedom of choice. The central issue is indeed that the choice of the individual is conditioned by the frame, which is an element external to her true self, the *homo œconomicus*.

We can now notice that a necessary condition for the efficiency of libertarian paternalism is that nudges effectively improve individual welfare, and therefore condition the choices of the individuals. If the choice architect is able to improve individual welfare thanks to

framing effects, then it means that choices are not free: they are indeed conditioned by an element external to the self, the will of the choice architect. The mere influence of nudges on individual choices implies that the freedom of choice is not preserved, since the individuals are to some extent "manipulated" by the choice architect. This means that the free-choice condition cannot be verified if a planner wants to improve individual welfare by using framing effects, as suggested by the proponents of libertarian paternalism. Libertarian paternalism, as conceived by Suntein and Thaler, cannot be libertarian by construction, since it relies on the idea that the choice architect can manipulate individual choices, which necessitates exploiting the limited freedom of choice of the individuals. Indeed, within Sunstein and Thaler's framework, the free-choice condition can be verified if and only if people are *homo œconomicus*: this would imply that they are not sensible to frames, and also that nudging is not necessary.

Libertarian paternalism presents several difficulties: the proponents of this approach firstly assume that there exists a true and rational self within each individual – similar to the paretian *homo œconomicus* – and that it constitutes a normative model of human behaviour. They then assume that it is possible to impartially isolate her true preferences, and therefore that the *homo œconomicus* provides an empirically relevant description of human behaviour in specific settings. It is also assumed that there exists a benevolent planner – who is an Econ and not a Human – who has the will and the ability to implement a choice architecture such that the individuals will *in fine* satisfy their true preferences. Finally, the freedom of choice is evaluated from the standpoint of individuals who are not subject to framing effects – i.e. Econs – and not from the standpoint of the

real Humans who make choices. The free-choice condition can therefore be verified if and only if the individuals are Econs, for whom nudges are by construction inefficient. In its current formulation, libertarian paternalism therefore constitutes a regular form of paternalism: the only difference between both approaches is that the planner does not directly coerce the individuals, but uses their limited freedom of choice to implement a specific option. It is therefore not possible to consider libertarian paternalism, as defined by Suntein and Thaler, as the "Real Third Way".

We now suggest reformulating libertarian paternalism and provide an alternative solution to the reconciliation problem. Our claim is that the difficulties of libertarian paternalism are due to a wrong diagnosis of the normative issue faced by boundedly rational individuals: in a libertarian perspective, the normative issue is not that the individuals do not necessarily satisfy their preferences, but that they are not autonomous, and therefore that a third party (the choice architect), thanks to framing effects, is able to manipulate them and to influence *in fine* their choices without their consent. Sunstein and Thaler justify paternalistic interventions by arguing that the individuals can make bad choices (they recognize it *a posteriori* and even agree in some cases to implement commitment devices to help them to achieve their objectives (2008, 44-49)) and are sensible to frames. We should therefore use frames to help them to correct their mistakes. The planner should then identify what they truly want and implement the adequate choice architecture. It is therefore necessary to define what a "bad" choice is and how to recognize it. The difficulty here is that we need to define a normative model of behaviour: a bad choice means indeed that there exists a difference between what an individual has chosen and

what she would have chosen if she was a “better person” (from her own point of view) at the moment of her choice. The notions of “mistake” and “bad” choice are indeed strongly related to the normative considerations of the individual and to her own perception as an agent. Consider for instance an individual who considers herself as the product of complex psychological phenomena. Unless her choices were influenced by external elements such as alcohol or drugs, she will not consider choices influenced by framing effects (resulting for instance from loss aversion) as mistakes or bad choices, since she made her choices as an autonomous agent. Libertarian paternalism, as formulated by Sunstein and Thaler, cannot be libertarian because it imposes to the individuals a normative model of behaviour, the paretian *homo œconomicus*.

However, although an individual, as an autonomous agent, can accept to be sensible to frames, being boundedly rational can constitute a normative issue since our choices can be manipulated by an external element to our self, the will of the choice architect. Providing a libertarian solution to the reconciliation problem therefore needs to directly treat the limited freedom of choice rather than its consequences (since the choice architect is unable to properly identify the self of each individual). Two possible approaches are then possible: the planner can either teach people how to make choices (rather than trying to guess what they would have chosen if they were able to make free choices), or ensures that the individual are able to implement their own commitment devices and the frames that will condition their choices. In the first case, the objective is to try to free the individuals from framing effects: a possible advice the planner could give to the individuals would be for instance to always evaluate the available options

under different frames. In the second case, we start from the observation that people (as psychological agents) are necessarily conditioned by frames, but that they should be able to design the frames of their future choices (rather than letting this task to a choice architect who will not necessarily try to get their consent): their freedom of choice is therefore preserved, since their choices are conditioned by a frame they have themselves designed as autonomous agents. Consider for instance the case of a doctor who must describe different possible treatments to a patient: rather than choosing the choice architecture that – according to her – will improve the well-being of the patient, she could for instance present a single information with different frames (for instance present the probability of success of each treatment and then the probability of failure, while emphasizing that the final choice can be influenced by the way with which the information was provided). Consider now the cafeteria of Sunstein and Thaler. Rather than directly choosing the location of the different items, Carolyn could inform the users how their choices can be affected, and then give them the opportunity to choose the location (by a public discussion and a vote for instance). If the users show little interest in this question and do not want to get involved in the choice of the location, they can delegate their choice to Carolyn and let her in charge of the location. The main difference between Sunstein and Thaler's example and this last situation is that, although Carolyn manipulates *in fine* the choice of the individual in both cases, she has their consent only in the second case, i.e. when the individuals, as autonomous agents, had the ability to choose the location but preferred to let a third party in charge of it.

The normative criterion we suggest is therefore not the maximization of one's self-assessed well-being any more, but the development of individual autonomy, understood as the ability to make free choices. The freedom of choice is evaluated as the autonomy towards external elements of the self, such as probably framing effects (either by learning, or by the possibility of choosing the frames of one's own choices): within Sunstein and Thaler's conception of the individual, the remaining criterion that determines the choices would therefore be the subjective evaluation of the different alternatives. This implies that individual autonomy is a sufficient condition to the maximisation of one's self-assessed well-being: instead of helping the individuals to maximize their welfare by nudging them towards what we think they would prefer if they were autonomous, we suggest helping the individuals to become autonomous, and then let them make their own choices. This new formulation of libertarian paternalism in terms of individual autonomy rather than preference-satisfaction implies that (i) the planner does not try to influence the choices of the individuals and therefore respects their subjectivity, and (ii) the individuals learn how to free themselves from the possible effects that could limit their freedom of choice. A major improvement of this reformulation is also that, unlike the proponents of the preference-satisfaction approach, we do not need to objectively define what the true self of the individuals is any more, and let them behave according to their own normative model of behaviour. It is therefore possible to define this approach as "soft" paternalism: the planner does not have an invasive role, and her action prevents the individuals from the "mistakes" that would have justified a paternalistic intervention.

5. Illustration: the management of a common-pool resource

We now suggest illustrating the relevance of a public policy guided by a principle of individual autonomy rather than preference-satisfaction by considering, on a more empirical level, the design principles that characterize robust institutions for managing common-pool resources (Ostrom 1990). We want to highlight that some of these principles are logical recommendations from the point of view of a government guided by a normative principle of individual autonomy.

Common-pool resources (CPR) are a class of goods characterized by two attributes, the difficulty of excluding individuals from benefiting from the resource, and the subtractability of the benefits consumed by an individual from those available to others. Two main types of problems can emerge in this context, appropriation and provision problems: appropriation problems are related to the exclusion of potential beneficiaries and the repartition of the output, whereas provision problems are related to the management of the stock of the resource, whether it be its creation, the maintenance or improvement of its production capabilities, or the avoidance of its destruction (Ostrom *et al*, 1994, 9). Ostrom (1990) suggested a list of eight design principles that characterize the institutions enabling a sustainable management of CPR, which have been slightly amended by Cox et al. (2010), who provide a meta-analysis of the different empirical works that tested those principles (extract from Cox *et al.* 2010, Table 4):

- **1A**, user boundaries: clear boundaries between legitimate users and non users must be clearly defined;
- **1B**, resource boundaries: clear boundaries are present that define a resource system and separate it from the larger biophysical environment;

- **2A**, congruence with local conditions: appropriation and provision rules are congruent with local social and environmental conditions;
- **2B**, appropriation and provision: the benefits obtained by users from a CPR, as determined by appropriation rules, are proportional to the amount of inputs required in the form of labour, material, or money, as determined by the provision rules;
- **3**, collective-choice arrangements: most individuals affected by the operational rules can participate in modifying the operational rules;
- **4A**, monitoring users: monitors who are accountable to the users monitor the appropriation and provision levels of the users;
- **4B**, monitoring the resource: monitors who are accountable to the users monitor the conditions of the resource;
- **5**, graduated sanctions: appropriators who violate operational rules are likely to be assessed graduated sanctions (depending on the seriousness and the context of the offense) by other appropriators, by officials accountable to the appropriators, or by both;
- **6**, conflict-resolution mechanisms: appropriators and their officials have rapid access to low-cost local arenas to resolve conflicts among appropriators or between appropriators and officials;
- **7**, minimal recognition of rights to organize: the rights of appropriators to devise their own institutions are not challenged by external governmental authorities;
- **8**, nested enterprises: appropriation, provision, monitoring, enforcement, conflict resolution, and governance activities are organized in multiple layers of nested enterprises

Our purpose is not to extensively discuss these different principles, but to highlight that most of them are directly supported by our normative criterion of individual autonomy.

We can indeed notice that the main feature of those principles is the idea that the users of the CPR should be able to design their own institutional environment (this is quite explicit in the principles 3 and 7). Furthermore, the eventual external actors who monitor the users and the resource, or who assess possible sanctions in case of non respect of the appropriation and provision rules are systematically accountable to the users. Several empirical studies showed for instance that when the rules are imposed by an external authority, this one generally fails to enforce them, leading to suboptimal results (Ostrom *et al.* 1994, 221-222). Nevertheless, although direct interventions often fail, the government can help the users to manage more efficiently the resource: Blomquist (1994) – from empirical evidence of groundwater systems in Southern California – suggests for instance that the design of provision and appropriation rules is facilitated by the presence of government agencies that can provide reliable information to the users (296-297). From various laboratory experiments and field studies, Ostrom *et al.* (1994) argue that the individuals can overcome the temptation of overusing the resource if they have some expectation of mutual trust, or the possibility of building trust through continued interaction and communication (328), and if they have some autonomy to decide on their own rules (323). However, since it appears that boundedly rational individuals can have some difficulties to reach optimal rules – mainly due to information issues and the complexity of the problem – governmental agencies play an important role by recognizing the right to the individuals to form their own rules and commitments, but also by providing them reliable information and backup enforcement mechanisms (322-327).

We can now notice that those conditions, and in particular the role of the government as an actor who provides information and support to the individuals without directly intervening nor trying to influence individuals' choices, correspond to the kind of normative prescriptions that would emerge from libertarian paternalism, when understood in terms of individual autonomy rather than preference-satisfaction. Our claim is indeed that the planner should assess public policies in terms of individual autonomy, i.e. the ability of the individuals to make free choices: this necessitates providing the largest information to the individuals, and let them decide on their own rules rather than imposing external rules.

In addition, the case of CPR gives us another argument in favour of a more deontological formulation of libertarian paternalism, the impact of institutional rules on individual preferences. It seems indeed that individual preferences in CPR situations depend on the institutional organisation that rules the appropriation and the provision of the resource: self-organized institutions are more likely to generate prosocial behaviours than rules imposed by an external authority. It means that – in addition of questioning the assumption of the existence of "true" preferences, independent from the institutional context – imposing the same policy can have a different impact according to its initiator: empirical evidence in CPR situations suggest that policies implemented by autonomous individuals are more likely to be efficient than policies implemented by an external choice architect. It is therefore probably not equivalent to try to implement what the individuals would have chosen if they were autonomous (such as in Sunstein and Thaler's definition of libertarian paternalism) and to try to directly improve the autonomy of the individuals. A nudge implemented by the individuals who will be affected by this nudge

is maybe more likely to be efficient than the same nudge implemented by an external choice architect: in the latter case, the individuals can indeed be suspicious about the objective of the choice architect, and then choose an option different from the one wanted by the choice architect.

The management of common-pool resources offers us a good illustration of one of the main objectives of our reformulation of libertarian paternalism. While the proponents of libertarian paternalism – as well as neoclassical economists – ground their normative assessments on consequentialist considerations such as the welfare generated by the satisfaction of one's preferences, we suggest adopting a more procedural approach by grounding our normative assessments on individual autonomy and the ability to design one's own frames. It seems indeed that the sustainable management of a CPR (and therefore the welfare it generates) is not only the result of the implementation of specific rules, but also of the conditions under which these rules were decided: promoting individual welfare therefore necessitates promoting individual autonomy, since the rules that will enable the individuals to maximize their welfare are more likely to be efficient if they are implemented by autonomous agents rather than by an external authority. If the objective of the government is the maximisation of individual self-assessed well-being, then a necessary preliminary step seems to be the development of individual autonomy.

References

Asch, S. E. 1955. "Opinions and Social Pressure". Scientific American, 193(5): 31-35.

Bleichrodt, H., J-L Pinto-Padres, and P. Wakker. 2001. "Making Descriptive Use of Prospect Theory to Improve the Prescriptive Use of Expected Utility". *Management Science*. 47: 1498-1514.

Blomquist, W. 1994. “Changing Rules, Changing Games: Evidence from Groundwater Systems in Southern California”. In *Rules, Games and Common-Pool Resources*, Ostrom, E., R. Gardner, and J. Walker. 293-300. Ann Arbor: University of Michigan Press.

Binmore, K. 1999. “Why Experiment in Economics?”. *Economic Journal.*109: F16-24.

Camerer, C. 2003. *Behavioral Game Theory*. Princeton: Princeton University Press.

Camerer, C., S. Issacharo, G. Loewenstein, T. O’Donaghue, and M. Rabin. 2003. “Regulation for Conservatives: Behavioral Economics and the Case for ‘Asymmetric Paternalism’”. Univ PA Law Rev, 151: 1211-1254.

Cox, M., G. Arnold, and S. Villamayor Tomas. 2010. “A Review of Design Principles for Community-based Natural Resource Management”. *Ecology and Society*. 15(4): 38.

Georgescu-Roegen, N. 1971. *The Entropy Law and the Economic Process*. Cambridge (Mass.): Harvard University Press.

Kahneman, D. 2011. *Thinking, Fast and Slow*. Allen Lane.

Kahneman, D. and A. Tversky. 1979. “Prospect Theory: an Analysis of Decision under Risk”. *Econometrica*, 47(2): 263-292.

Kahneman, D. and A. Tversky. ed. 2000. *Choice, Value, and Frames*. Cambridge: Cambridge University Press.

Kahneman, D., P. Wakker and R. Sarin. 1997. “Back to Bentham? Explorations of Experienced Utility”. *Quarterly Journal of Economics*, 112: 375-405.

Kavka, G. 1983. “The Toxin Puzzle”. *Analysis*. 43.

McQuillin, B. and R. Sugden. 2012. “Reconciling Normative and Behavioural Economics: the Problems to be Solved”. *Social Choice and Welfare*, 38(4): 553-567.

Milgram, S. 1975. Obedience to Authority. New York: Harper Colophon.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action.* New-York: Cambridge University Press.

Ostrom, E., R. Gardner, and J. Walker. 1994. *Rules, Games and Common-Pool Resources*. Ann Arbor: University of Michigan Press.

Pareto, V. 1936 [1916]. *The mind and society: a treatise on general sociology*, translated by Andrew Bongiorno and Arthur Livingston. London: Jonathan Cape.

Pareto, V. 1971 [1909]. *Manual of Political Economy*, translated by A. Schwier from the 1927 french edition *Manuel d'économie politique* (Genève: Droz). London: McMillan

Plott, C. 1996. “Rational individual behaviour in markets and social choice processes: the discovered preference hypothesis”. In *The rational foundations of economic behaviour*, ed. K.J. Arrow, E. Colombatto, M. Perlman and C. Schmidt, 225-50. International Economic Association and Macmillan.

Samuelson, W. and R.J. Zeckhauser. 1988. “Status Quo Bias in Decision Making”. *Journal of Risk and Uncertainty*. 1: 7-59.

Sugden, R. 2004. “The Opportunity Criterion: Consumer Sovereignty Without the Assumption of Coherent Preferences”. *American Economic Review*, 94: 1014-1033.

Sugden, R. 2007. “The Value of Opportunities Over Time When Preferences are Unstable”. *Social Choice and Welfare*, 29: 665-682.

Sunstein, C. 1998. “Selective Fatalism”. *The Journal of Legal Studies*. 27(S2): 799-823.

Sunstein, C. and R. Thaler. 2003. “Libertarian Paternalism is not an oxymoron". *Univ Chic Law Rev*, 70: 1159-1202.

Thaler, R. and C. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, New Haven.